EigenNet in High Dimensional Space

Feature Selection Method

Rahul Barikeri, Ashutosh Modi, Rubein Shaikh

University of South Florida

December 5, 2016

Outline

1. Introduction
2. Feature Selection
3. EigenNet
4. Results

Feature

Table 1.2 Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

[1]

[1] Data Mining: Practical Machine Learning Tools and Techniques: Ian H. Witten, Eibe Frank, Mark A. Hall

Feature Selection Vs Extraction

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \rightarrow \begin{bmatrix} X_a \\ X_b \\ \vdots \\ X_j \end{bmatrix} \qquad f\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}\right) \rightarrow \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_k \end{bmatrix}$$

Huges phenomenon

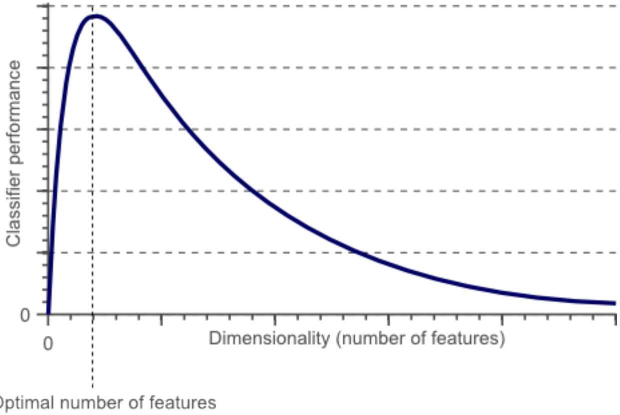
- Curse of Dimensionality[2]
- Enough examples are required to have sufficient combination of features

Challenge

For small number of instances, prediction accuracy decreases with increase in dimensionality

[2] Frame Hughes, Gordon. "On the mean accuracy of statistical pattern recognizers." IEEE transactions on information theory 14.1 (1968): 55-63.

Performance Curve



3

3 http://www.visiondummy.com/wp-content/uploads/2014/04/dimensionality_vs_performance.png

Feature Selection Types

- Wrapper
- Filter
- Embedded

Feature Selection Types

- Wrapper
  - 2^N combinations
- Filter
- Embedded

Feature Selection Types

- Wrapper
  - 2^N combinations
- Filter
  - Variable ranking technique
- Embedded

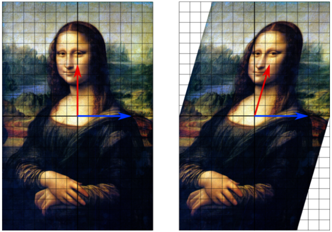
Feature Selection Types

- Wrapper
  - 2^N combinations
- Filter
  - Variable ranking technique
- Embedded
  - Part of training algorithm

Eigen values and Eigen Vectors

$$A\overrightarrow{V} = \lambda \overrightarrow{V}$$

A = Transformation matrix
$\overrightarrow{V}$ = Eigen Vector
λ = Eigen Value



4

[4] https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

Classification Objective

Cost function

$$\min_{\beta} J(\beta) = \frac{1}{2N} \sum_{i=1}^{N} \left(f_{\beta}\left(x^{(i)}\right) - y^{(i)} \right)^2$$

For Logistic regression

$$f_{\beta}(x) = \frac{1}{1 + e^{-\beta^T x}}$$

Modified Classification Objective

With constraint

$$\min_{\beta\,\lambda} J(\beta) = \frac{1}{2N}\sum_{i=1}^{N}\left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\|\beta\|_p$$

p norm:

$$\|\beta\|_p = \left(\sum_{i=1}^{N}|\beta|^p\right)^{\frac{1}{p}}$$

l_0 Norm : Pseudo Norm (Hard to optimize)
l_1 Norm : LASSO
l_2 Norm : Ridge Regression

Why regularization?

- To avoid overfitting

Why regularization?

- To avoid overfitting
- To put constraint on values of weight

Why regularization?

- To avoid overfitting
- To put constraint on values of weight
  - Weight Decay or Parameter Shrinkage

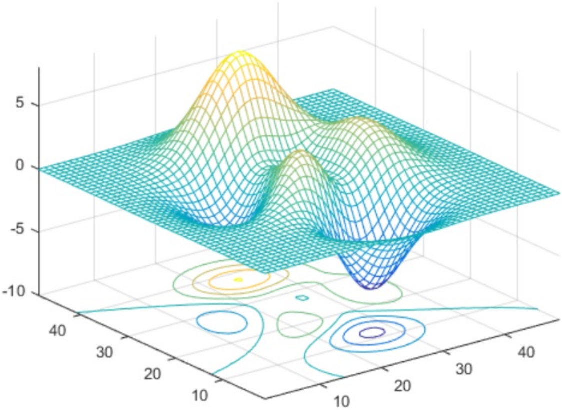
Why regularization?

- To avoid overfitting
- To put constraint on values of weight
  - Weight Decay or Parameter Shrinkage
- Allows model to be trained on limited data sets.

Why regularization?

- To avoid overfitting
- To put constraint on values of weight
  - Weight Decay or Parameter Shrinkage
- Allows model to be trained on limited data sets.
- Improves prediction for new instance

Gradient Descent and Contours

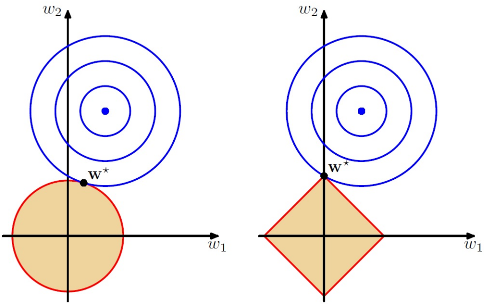


5

[5] [X,Y] = meshgrid(-3:.125:3); Z = peaks(X,Y); meshc(Z)

L1 and L2 Norm

Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^{\star}$. The lasso gives a sparse solution in which $w_1^{\star} = 0$.



[6]

[6] Pattern Recognition and Machine Learning: Christopher M. Bishop

LASSO

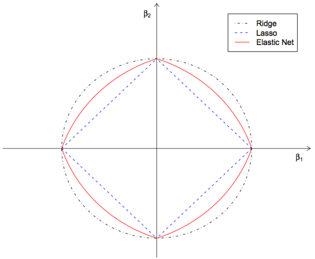
- Least Absolute Selection and Shrinkage Operator
- L1 norm
- Advantages
  - For high dimension space coefficients tend to 0
  - Sparsification of $\overrightarrow{\beta}$
  - Which feature has real impact select them
- Feature slection in linear model
- Disadvantages
  - Fails to perform group selection
  - Total # of feature selection is bounded by # of instances

Ridge Regression

- L2 norm
- Advantages
  - For correlated features puts constraint on their weights
- Disadvantages
  - Can not perform feature selection unless $\lambda \to \infty$

Elastic Net

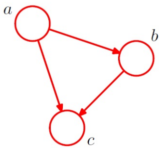
- For high dimensional data correlation between features can be high
- Correlated variable sometimes form a group
- If one is selected from this group we would want to select entire group
- Uses both l_1 and l_2 norm



Graphical Model

- Predict multiple variables that depend on each other
- Graphical models model dependencies between output variables
- We want to predict an output vector $\mathbf{y}$ of random variables given an observed feature vector $\mathbf{x}$

$$p(a, b, c) = p(c|a, b)\, p(b|a)\, p(a)$$



7

[7] Pattern Recognition and Machine Learning: Christopher M. Bishop

Bayesian Approach

Goal is to compute posterior over model

Data Likelihood

$$p\left(y|X,\, w\right) = \prod_{i=1}^{n} \sigma\left(y_i w^T x_i\right)$$

where,

$$\sigma\left(z\right) \equiv \textit{Gaussian Cumulative Distribution}$$

For Laplace prior distribution elastic net regularizer

$$p\left(w\right) = \prod_{j} \exp\left(-\lambda_1 |w_j| - \lambda_2 w_j^2\right)$$

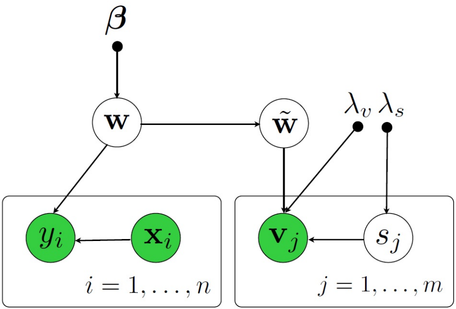
Generative and Discriminative

- Generative
  - Able to generate synthetic data points
  - e.g. Gaussian mixture model, Naive Bayes
- Discriminative
  - Uses conditional probability distribution
  - e.g. Logistic regression, Support Vectore Machines

EigenNet

- Is hybrid of conditional and generative model
- Conditional model
    - Learns classifier $\rightarrow$ selecting eigen vectors
- Generative model
    - captures correlation $\rightarrow$ estimation
- Eigenstructure guiding variable selection

EigenNet

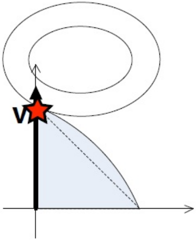


$\lambda_v \equiv$ Hyperparameter
$v_j \equiv$ Eigen Vector
$s_j \equiv$ Scaling factor

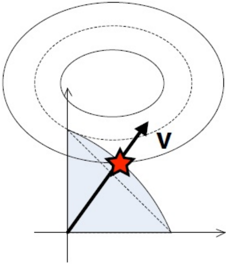
8

[8] Qi, Yuan, and Feng Yan. EigenNet: A Bayesian hybrid of generative and conditional models for sparse learning. arXiv preprint arXiv:1102.0836 (2011).

EigenNet



When variables are uncorrelated



When variables are correlated

9

[9] Dr. Alan Qi: Purdue CS & Statistics

Data set

Microarray Datasets

	Leukemia (ALL-AML)	Ovarian
Number of Attributes	7129	15154
Number of Instances	72	253
Classes	47 (ALL)	91 (Normal)
	25 (AML)	162 (Cancer)

[10]

- ALL : Acute Lymphoblastic Leukemia
- AML: Acute Myeloblastic Leukemia

[10] Dataset: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

Observations

- AUC and Accuracy

	Ovarian	Leukemia
Decision Tree	0.957 (95%)	0.784 (83%)
LibSVM	0.819 (87%)	0.5 (65%)
ElasticNet	1 (100%)	0.97 (96%)

Observations

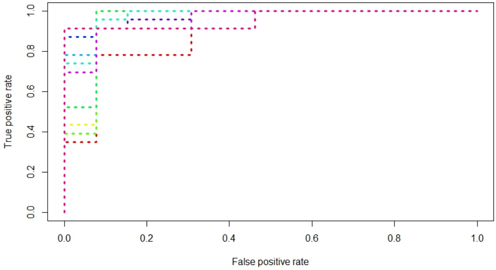


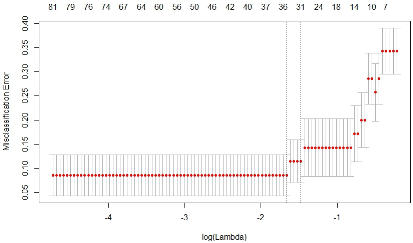
Figure: Leukemia ROC curve

11

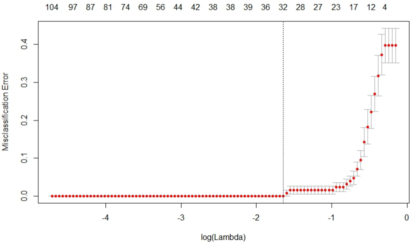
[11] RStudio: Library - pROC

Observations

- CV curve



(a) Leukemia



(b) Ovarian

12

[12] RStudio: glmnet - plot()

Thank you

Questions!